

Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English

Hermann Moisl (hermann.moisl@ncl.ac.uk)

Warren Maguire (w.n.maguire@ncl.ac.uk)

Will Allen (w.h.a.allen@ncl.ac.uk)

Abstract

The newly-created Newcastle Electronic Corpus of Tyneside English (NECTE) offers an opportunity to study a recent sample of English spoken in the Tyneside region of North-East England. This paper describes an exploratory multivariate analysis of phonetic data derived from NECTE that was undertaken with the aim of generating hypotheses about phonetic variation among speakers and speaker groups in the corpus, and how this variation correlates with social factors. The discussion is in four main parts. The first part outlines exploratory multivariate analysis in general and hierarchical cluster analysis in particular, the second describes the NECTE phonetic data used in the analysis, the third carries out a hierarchical cluster analysis of that data, and the fourth interprets the cluster analysis and relates the result to existing work on Tyneside English. The interpretation of the cluster analysis result is that phonetic variation among the NECTE speakers correlates strongly with gender and to a lesser extent with socio-economic status, but a correlation with age could not be demonstrated. The conclusion, finally, indicates directions for future work.

*

Keywords

Data preprocessing; data representation; dialectology; dimensionality reduction; document length normalization; exploratory multivariate analysis; hierarchical cluster analysis; hypothesis generation; phonetic analysis; sociolinguistics; Tyneside English.

*

The newly-created Newcastle Electronic Corpus of Tyneside English (NECTE) offers an opportunity to study a recent sample of English spoken in the Tyneside region of North-East England. This paper describes an exploratory multivariate analysis of phonetic data derived from NECTE that was undertaken with the aim of generating hypotheses about phonetic variation among speakers and speaker groups in the corpus, and how this variation correlates with social factors.

The discussion is in four main parts. The first part outlines exploratory multivariate analysis in general and hierarchical cluster analysis in particular, the second describes the NECTE phonetic data used in the analysis, the third carries out a hierarchical cluster analysis of that data, and the fourth interprets the cluster analysis and relates the result to existing work on Tyneside English. The conclusion indicates directions for future work.

1. Exploratory multivariate analysis

a) Introduction to multivariate analysis

The proliferation of computational technology has generated an explosive production of electronically encoded information of all kinds. In the face of this, traditional paper-based methods for search and interpretation of data have been overwhelmed by sheer volume, and a wide variety of computational methods has been developed in an attempt to make the deluge at least tractable. As such methods have been refined and new ones introduced, something over and above tractability has emerged – new and unexpected ways of understanding the data. The fact that a computer can deal with vastly larger data sets than a human is an obvious factor, but there are two others of at least equal importance: one is the ease with which data can be manipulated and reanalyzed in interesting ways without the often prohibitive labour that this would involve using manual techniques, and the other is the extensive scope for visualization that computer graphics provide.

These developments have clear implications for the analysis of large bodies of text in corpus-based linguistics. On the one hand, large electronic text

corpora potentially exploitable by the linguist are being generated as a by-product of the many kinds of daily IT-based activity worldwide, and, on the other, more and more application-specific electronic linguistic corpora are being constructed. Effective analysis of such corpora will increasingly be tractable only by adapting the interpretative methods developed by the statistical, information retrieval, and related communities (Tabachnik / Fidell 2001, Hair *et al.* 1998, Baeza-Yeates / Ribeiro-Neto 1999, Belew 2000). In the present paper we are interested in one particular type of tool: multivariate analysis. What is multivariate analysis?

Observation of nature plays a fundamental role in science. In current scientific methodology, an hypothesis about some natural phenomenon is proposed and its adequacy assessed using data obtained from observation of the domain of inquiry. But nature is dauntingly complex, and there is no practical or indeed theoretical hope of being able to observe even a small part of it exhaustively. Instead, the researcher selects particular aspects of the domain for observation. Each selected aspect is represented by a variable, and a series of observations is conducted in which, at each observation, the values for each variable are recorded. A body of data is thereby built up on the basis of which an hypothesis can be assessed. One might choose to observe only one aspect – the height of individuals in a population, for example-- in which case the data set consists of more or less numerous values assigned to one variable; such a data set is referred to as univariate. If two values are observed – say height and weight— then the data set is said to be bivariate, if three --height, weight, age-- trivariate, and so on up to some arbitrary number n . Strictly speaking, any data set where n is greater than 1 is multivariate, though in practice that term is normally used only when n is greater than 2 or 3 (Hair *et al.* 1998, Tabachnik / Fidell 2001).

As the number of variables grows, so does the difficulty of understanding the data, that is, of conceptualizing the interrelationships of variables within a single data item on the one hand, and the interrelationships of complete data items on the other. Multivariate analysis is the computational use of

mathematical and statistical tools for understanding these interrelationships in data.

Numerous techniques for multivariate analysis exist. They can be divided into two main categories which are usually referred to as 'exploratory' and 'confirmatory'. Exploratory analysis aims to discover regularities in data which can serve as the basis for formulation of hypotheses about the domain from which the data comes; such techniques emphasize intuitively accessible, usually graphical representations of data structure. Confirmatory analysis attempts to determine whether or not there are significant relationships between some number of selected independent variables and one or more dependent ones. These two types are complementary in that the first generates hypotheses about data, and the second tries to determine whether or not the hypotheses are valid. Exploratory analysis is naturally prior to confirmatory; this discussion is concerned with the former.

b) Hierarchical cluster analysis

Hierarchical cluster analysis is a variety of exploratory multivariate analysis. To understand how it works and how the results it gives should be interpreted, it is first necessary to understand the concept of distance between data points in vector space.

Assume a domain of inquiry, say a linguistic corpus, which will be studied using six variables. If the six-dimensional data is to be analyzed using an exploratory method, it has to be represented mathematically. This is done in the form of vectors, where a vector is a sequence of values indexed by the positive integers 1, 2, 3.... Thus, figure 1

$$v = \begin{array}{|c|c|c|c|c|c|} \hline 1.6 & 2.4 & 7.5 & 0.6 & 0.1 & 2.6 \\ \hline 1 & 2 & 3 & 4 & 5 & 6 \\ \hline \end{array}$$

Figure 1: Example of a vector

is a length-6 vector v of numerical values in which the value of v_1 is 1.6, the value of v_2 is 2.4 and so on. Where the data consists of more than one case, which it usually does, then each case is represented by a vector, and the set

of vectors is assembled into a matrix, which is a sequence of vectors arranged in rows and the rows are indexed by the positive integers 1, 2, 3... . In matrix M , case 2 is at row M_2 and the value of the third variable for that case is at $M_{2,3}$, that is, 0.1.

$$M = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 1.6 & 2.4 & 7.5 & 0.6 & 0.1 & 2.6 \\ 2 & 3.4 & 6.2 & 0.1 & 1.1 & 0.1 & 1.1 \\ 3 & 10 & 9.1 & 9.0 & 5.2 & 9.0 & 5.2 \\ \hline & 1 & 2 & 3 & 4 & 5 & 6 \end{array}$$

Figure 2: Example of a matrix

A vector space is a geometrical interpretation of a set of vectors:

- The dimensionality n of the vectors, that is, the number of its elements, defines an n -dimensional space
- The indices of the vectors define the coordinates of the space
- The values in the vector define the coordinates of a point in that space

For example, a bivariate data set defines a 2-dimensional space in which each vector specifies the coordinates of a point in that space. Take a data set consisting of vectors that specify the age and weight of some number of individuals. A single such vector might be $v = (36, 160)$. In geometrical terms, the x or age axis is 0..100, the y or weight axis is 0..200, and any vector in the data set can be plotted in the (x, y) space, as in Figure 3:

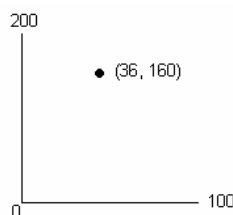


Figure 3: A vector in 2-dimensional space

If more vectors are plotted in the space, nonrandom structure may or may not emerge, depending on the interrelationships of the real-world characteristics that the variables represent. Where there are no structured real-world interrelationships, the result will look something like the plot of random points

in figure 4a. If there is structure, the plot might look something like 4b, where two clusters have clearly emerged. These clusters tell us something about the interrelationships of the represented entities.

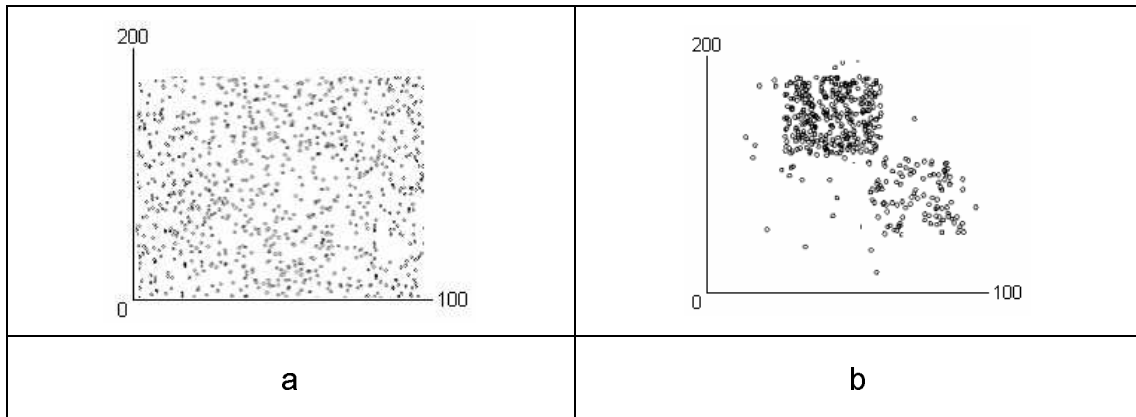


Figure 4: Plots of random and nonrandom data vectors

Analogously, a trivariate (age, weight, height) vector $v = (36, 160, 71)$ from a data set of length-3 vectors defines a point in 3-dimensional space, as shown in Figure 5:

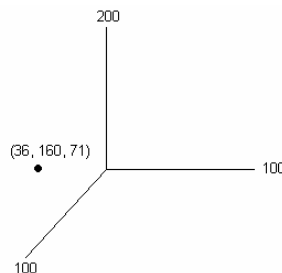


Figure 5: A vector in 3-dimensional space

A length-4 vector defines a point in 4-dimensional space, and so on to any dimensionality n . Mathematically there is no problem with spaces of dimension greater than 3: the conceptual and formal frameworks apply to n -dimensional spaces, for any n , as straightforwardly as to 2 or 3 dimensional ones. The only problems lie in the possibility of visualization and intuitive understanding. As the number of variables, and thus dimensions, grows beyond 3, graphical representation and intuitive comprehension of it become impossible -- who can visualize points in a four-dimensional space, not to speak of a 40-dimensional one?

Given that the structure of data with dimensionality higher than 3 cannot be directly visualized, how is it to be understood? The various exploratory multivariate methods provide indirect visualizations. Hierarchical cluster analysis, in particular, constructs ‘ dendrograms’ or trees that show the constituency structure of clusters using relative distance between and among points in the high-dimensional data vector space, where ‘ distance’ can for present purposes be understood quite literally: distance between points A and B in figure 6 can be measured, and it is less than the measured distance between, say, A and C.

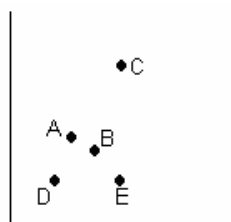


Figure 6: Vectors with various relative distances in 2-dimensional space

These relativities can be represented as a tree in which the horizontal lines represent distance: the longer the line, the greater the distance. Knowing this, it is easily seen that, in figure 7, there are two main clusters, (C) and (ABDE), the latter of which itself has internal structure.

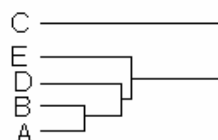


Figure 7: Tree representation of relative vector distances in figure 6

Given a set of data vectors, hierarchical cluster analysis generates the corresponding distance-based tree. Details of distance measures and how these are used to generate such cluster trees are available in a wide variety of textbooks – for example Everitt (2001).

2. The NECTE data

a) Overview of the NECTE corpus

The NECTE corpus is based on two pre-existing corpora of audio-recorded speech, one of them gathered by the Tyneside Linguistic Survey (TLS) undertaken in the late 1960s (Strang 1968, Pellowe *et al.* 1972, Pellowe / Jones 1978, Jones-Sargent 1983), and the other between 1991 and 1994 for the 'Phonological Variation and Change in Contemporary Spoken English' (PVC) project (Milroy *et al.* 1994, Docherty / Foulkes 1999). The aim of the NECTE project has been to enhance, improve access to, and promote the re-use of the TLS and PVC corpora by amalgamating them into a single, TEI-conformant electronic corpus. The result is now available to the research community in a variety of formats: digitized sound, phonetic transcription, and standard orthographic transcription, all aligned and accessible on the Web (Corrigan / Moisl / Beal 2005).

The TLS component of NECTE includes phonetic transcriptions of about 10 minutes of each of 63 recordings. It is with these transcriptions that the remainder of this discussion is concerned.

b) The TLS phonetic transcriptions

One of the main aims of the TLS project was to see whether systematic phonetic variation among Tyneside speakers of the period could be significantly correlated with variation in their social characteristics. To this end they developed a methodology which was radical at the time and remains so today: in contrast to the then-universal and still-dominant theory driven approach, where social and linguistic factors are selected by the analyst on the basis of some combination of an independently-specified theoretical framework, existing case studies, and personal experience of the domain of enquiry, the TLS proposed a fundamentally empirical approach in which salient factors are extracted from the data itself and then serve as the basis for model construction.

To realize its research aim using its empirical methodology, the TLS had to compare the audio interviews it had collected at the phonetic level of

representation. This required that the analog speech signal be discretized into phonetic segment sequences, or, in other words, to be phonetically transcribed. Details of the TLS transcription scheme are available in Jones-Sargent (1983) and Corrigan / Moisl / Beal (2005). For present purposes, it is sufficient to note that two levels of transcription were produced, a highly detailed narrow one designated 'State', and a superordinate ' Putative Diasystemic Variables' (PDV) level which collapsed some of the finer distinctions transcribed at the ' State' level.

3. Hierarchical cluster analysis of the TLS phonetic transcriptions

This section applies hierarchical cluster analysis to the TLS phonetic transcriptions at the PDV level of phonetic representation.

a) Data construction

The analyses are based on comparison of profiles associated with each of the TLS speakers. A profile for any speaker S is the number of times S uses each of the PDV codes defined by the TLS transcription scheme in his or her interview. More specifically, the profile P associated with S is a vector having as many elements as there are codes such that each vector element P_j represents the j ' th PDV, where j is in the range 1..number of codes in the TLS scheme, and the value stored at P_j is an integer representing the number of times S uses the j ' th PDV code. There are 156 PDVs, and so a PDV profile is a length-156 vector.

There are 63 TLS speakers, and their profiles are represented in a matrix having 63 rows, one for each profile. At the PDV level, therefore, the data used in this study is a 63 x 156 matrix M.

b) Data preprocessing

Prior to analysis, M was transformed in two ways.

i. Normalization for text length

The number of codes per transcription varies significantly. This variation in length has to be taken into account when conducting the analyses in order to avoid skewed results. The following function was applied to the raw PDV frequency matrix M:

$$freq'(M_{ij}) = freq(M_{ij}) \times \left(\frac{\mu}{l} \right)$$

Figure 8: Text length normalization function

where $freq'$ is the adjusted frequency, M_{ij} is the value at the (i,j) coordinates of the data matrix M, $freq$ is the raw frequency, μ is the mean number of codes per interview across all 63 interviews, and l is the number of codes in interview i . This function increases the frequency values for relatively shorter interviews in proportion to the mean interview length, and decreases frequency values for relatively longer interviews relative to the mean.

ii. Dimensionality reduction

Since there are 156 PDVs, there are 156 criteria for distinguishing the 63 speakers. It is, however, easy to show that many of these criteria are superfluous. The key to doing so is the statistical concept of variance, which measures the range of variation of values that a variable takes. Each of the columns in M is a variable; the variances of the columns were calculated, sorted by decreasing magnitude, and plotted, and the result is shown in figure 9:

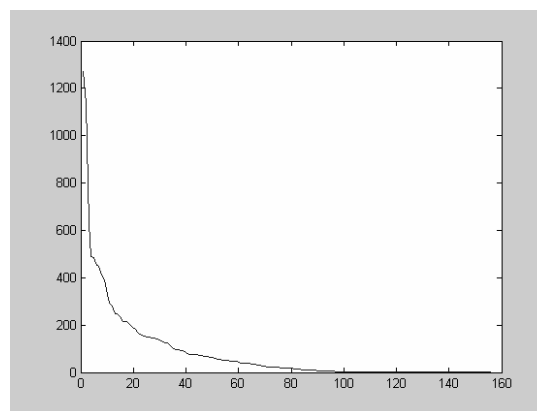


Figure 9: Variance profile of the 156 PDV variables in the data matrix

M

Relative to the variance range 0..1300 there are a few very high-variance PDVs, a moderate number of middling-variance PDVs, and a majority of low-variance ones. Low variance says that the values a variable takes do not vary a great deal. This makes them unimportant for distinguishing the cases that the variable describes. In the present case, the PDVs to the right of – generously—the 80th have such low variance that they can be eliminated from consideration. They were, therefore, removed from M, resulting in a reduced-dimensionality 63 x 80 matrix.

c) Data analysis

Hierarchical cluster analysis is not a single method but a family of closely related ones which offer a range of ways to measure distance between vectors in n -dimensional space and of defining what constitutes a cluster in terms of those distance measures. Details can be found in any multivariate analysis or cluster analysis textbook; a standard account is in Everitt (2001). M is here analyzed using one particular combination of distance measure and cluster definition: squared Euclidean distance and increase in sum of squares clustering, also known as Ward's Method. This combination was chosen to facilitate comparison with earlier work on the data being analyzed here, on which more later. A standard caution in hierarchical cluster analysis applies, however. Relative to a given data matrix, different distance measure / cluster definition combinations can and usually do generate different trees. This leads to an obvious question: what are these methods really telling us about the structure of the data they describe --how reliable, in other words, are they, and are they in fact any use at all if they cannot be relied on to reveal the true structure of the data? The answer is that this is the wrong way of looking at what these methods are useful for. A set of vectors in n -dimensional space has a 'true' structure in the sense that the relative distances among the vectors exist independently of the observer and can be determined to arbitrary accuracy using an appropriate measure. Cluster membership, however, is not latent in the data. It is a matter of definition: each clustering method defines a cluster in its own way and then describes the data in terms of that definition, giving its own characteristic view of it. When faced with different analyses of

the same data, it is up to the analyst to understand their artifactual nature, to realize that none is necessarily more 'true' of the data than any other, and to select those that are most useful for hypothesis generation, which is the object of the exploratory multivariate analysis.

Two analyses of the PDV frequency data are presented: the first (figure 10a) includes all 63 speakers, that is, 7 Newcastle and 56 Gateshead speakers in the sample, and the second (figure 10b) looks in detail at the Gateshead speakers.

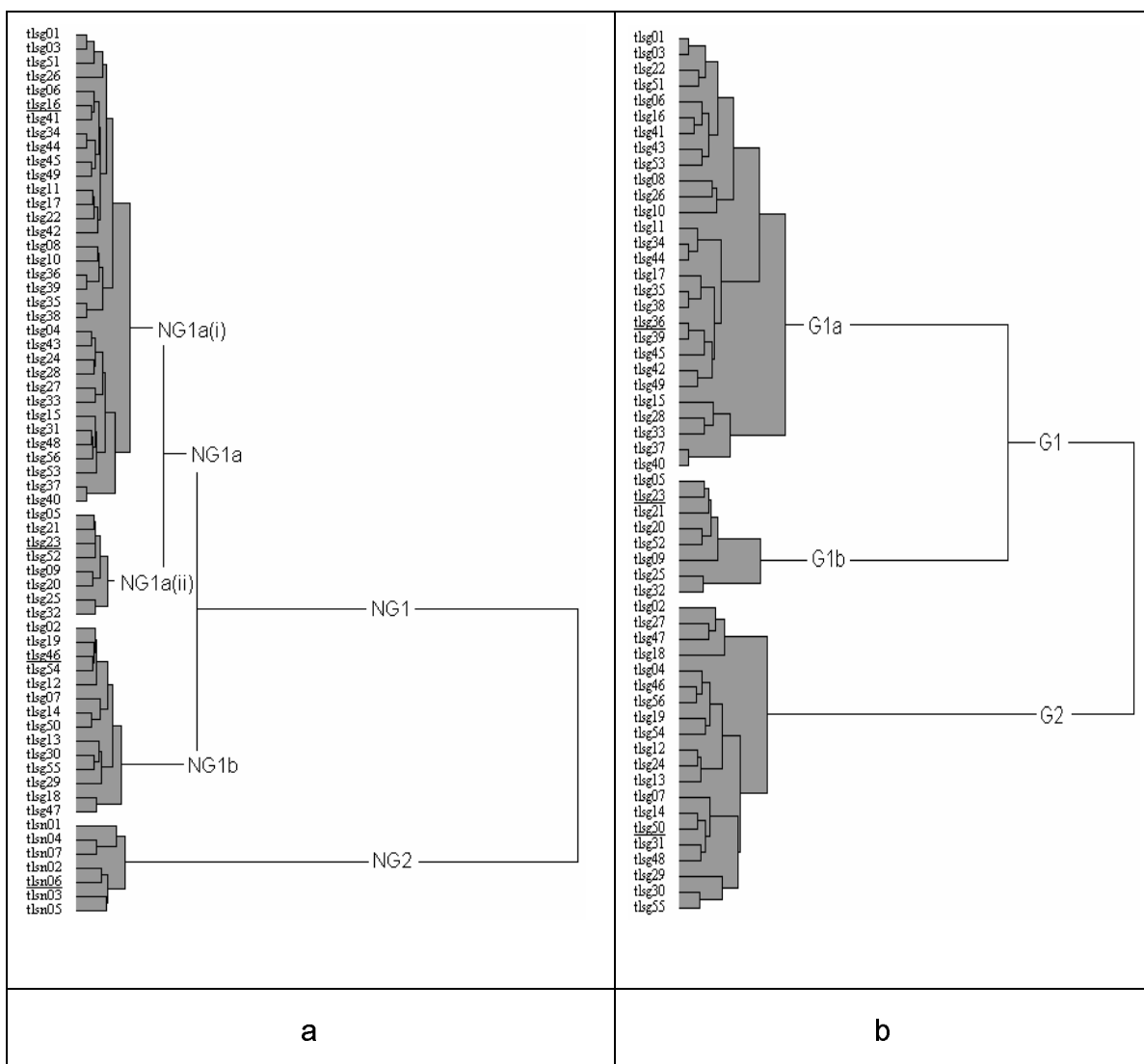


Figure 10: Cluster trees of the NECTE data matrix

Analysis 1: All speakers

There are two main clusters, labelled NG1 and NG2, and NG1 has subclusters with labels indicating constituency. NG2 clusters markedly against the rest, and comprises the Newcastle group. On the basis of the phonetic segment frequency distribution evidence at the PDV level, therefore, Newcastle speakers are strongly distinguished from Gateshead ones.

Analysis 2: Gateshead speakers only

The purpose of this second analysis is to examine in detail the structure of the cluster of Gateshead in Analysis 1, and to see if that structure correlates interestingly with social characteristics such as gender, age, and socio-economic status of the TLS speakers. We were primarily interested in the vowel variables, and so looked only at the vowel PDVs, though a similar analysis could be done for the consonants. The PDV frequency matrix M was re-calculated using vowel-PDV frequency data for the Gateshead speakers only, length-normalized, and dimensionality reduced as above to a 56 x 40 matrix. The result was two main clusters, labelled G1 and G2, and G1 itself comprises two subclusters G1a and G1b.

In the following table, the cluster labels are given in the first column, followed by the NECTE speaker ID and social variables selected from those included in the NECTE corpus.

Cluster	ID	Sex	Age	Education	Employment	Cluster	ID	Sex	Age	Education	Employment
G1a	tsg01	F	31-40	Minimum	Skilled manual	G1b	tsg05	F	21-30	Day release	Lower admin
G1a	tsg03	F	41-50	Minimum	Semi-skilled manual	G1b	tsg23	M	21-30	Tertiary	Higher admin
G1a	tsg22	F	41-50	Minimum	Skilled manual	G1b	tsg21	F	17-20	Night school	Skilled manual
G1a	tsg51	F	21-30	Minimum	Skilled manual	G1b	tsg20	M	21-30	Tertiary	Skilled manual
G1a	tsg06	F	61-70	Minimum	Semi-skilled manual	G1b	tsg52	F	31-40	Night school	Skilled manual
G1a	tsg16	F	41-50	Minimum	Unskilled manual	G1b	tsg09	F	21-30	Day release	Higher admin
G1a	tsg41	F	51-60	Minimum	Unskilled manual	G1b	tsg25	F	41-50	Night school	Skilled manual
G1a	tsg43	M	21-30	Minimum	Skilled manual	G1b	tsg32	F	51-60	Minimum	Lower admin
G1a	tsg53	M	16-20	Day release	Skilled manual						
G1a	tsg08	F	17-20	Minimum	Unskilled manual	G2	tsg02	M	31-40	Minimum	Skilled manual
G1a	tsg26	F	41-50	Minimum	Unskilled manual	G2	tsg27	M	21-30	Minimum	Skilled manual
G1a	tsg10	F	17-20	Minimum	Unskilled manual	G2	tsg47	M	21-30	Minimum	Unskilled

											manual
G1a	tsg11	F	31-40	College	Semi-skilled manual	G2	tsg18	M	31-40	Minimum	Unskilled manual
G1a	tsg34	F	31-40	College	Lower admin	G2	tsg04	M	61-70	Minimum	Skilled manual
G1a	tsg44	F	51-60	Night school	Unskilled manual	G2	tsg46	M	31-40	Minimum	Lower admin
G1a	tsg17	F	51-60	Minimum	Semi-skilled manual	G2	tsg56	M	21-30	Night school	Skilled manual
G1a	tsg35	F	31-40	Minimum	Unskilled manual	G2	tsg19	M	41-50	Minimum	Semi-skilled manual
G1a	tsg38	F	31-40	Minimum	Unskilled manual	G2	tsg54	M	21-30	Day release	Skilled manual
G1a	tsg36	F	31-40	Minimum	Unskilled manual	G2	tsg12	M	21-30	Minimum	Semi-skilled manual
G1a	tsg39	F	31-40	Minimum	Unskilled manual	G2	tsg24	M	61-70	Minimum	Skilled manual
G1a	tsg45	F	41-50	Minimum	Unskilled manual	G2	tsg13	M	61-70	Minimum	Unskilled manual
G1a	tsg42	F	21-30	Minimum	Skilled manual	G2	tsg07	M	31-40	Minimum	Skilled manual
G1a	tsg49	F	41-50	Minimum	Unskilled manual	G2	tsg14	M	41-50	Minimum	Semi-skilled manual
G1a	tsg15	F	21-30	Minimum	Skilled manual	G2	tsg50	M	31-40	Minimum	Skilled manual
G1a	tsg28	M	61-70	Minimum	Unskilled manual	G2	tsg31	M	31-40	Minimum	Unskilled manual
G1a	tsg33	M	61-70	Minimum	Unskilled manual	G2	tsg48	M	51-60	Night school	Skilled manual
G1a	tsg37	F	41-50	Minimum	Unskilled manual	G2	tsg29	M	41-50	Minimum	Skilled manual
G1a	tsg40	F	41-50	Minimum	Unskilled manual	G2	tsg30	M	41-50	Minimum	Skilled manual
						G2	tsg55	M	21-30	Minimum	Skilled manual

Figure 11: Correlation of clusters with social variables

The clearest correlation is between cluster structure and sex: G2 consists entirely of men, and G1 mainly though not exclusively of women. With a few slight exceptions, the men in G2 have the minimum legal level of education, and all are in unskilled - skilled manual employment. In G1 there is a clear split between a cluster consisting mainly of women with minimum education in unskilled - skilled manual employment, and one consisting of men and women with a slightly higher educational and employment level. Finally, there is no obvious correlation between cluster structure and age.

b) Assessment of results relative to existing work on Tyneside English

The TLS project culminated in Jones-Sargent (1983), who performed cluster analyses based on the segmental phonological data and on the social data we have been dealing with, and then attempted to relate the two in a

sociolinguistically meaningful way. In order to derive social and linguistic classifications, Jones-Sargent used hierarchical cluster analysis, and squared Euclidean distance and Ward' s method more specifically. We chose that combination in our own analyses to facilitate comparison with Jones-Sargent' s work. Direct comparison is nevertheless complicated, for several methodological reasons:

- The data that Jones-Sargent used for one speaker (in her analysis, labeled STEPH5M3) is no longer directly available, while data for 12 speakers (tlsg31-tlsg40, tlsg55, tlsg56) included in the present analysis were not analyzed by Jones-Sargent.
- Jones-Sargent analyzed the detailed State rather than the PDV level used in the present study.
- A different length normalization method was used (Jones-Sargent 1983:93-4)
- The data was not dimensionality-reduced.
- Due to computational hardware and software limitations when the analysis was done in the early 1980s, the TLS data had to be partitioned into three groups --monophthongs, diphthongs, and consonants-- and analyzed separately (Jones-Sargent 1983:105 & 195-199), and the cluster trees for the groups differ among themselves.

Jones-Sargent's tree for the diphthong group is given in figure 12:

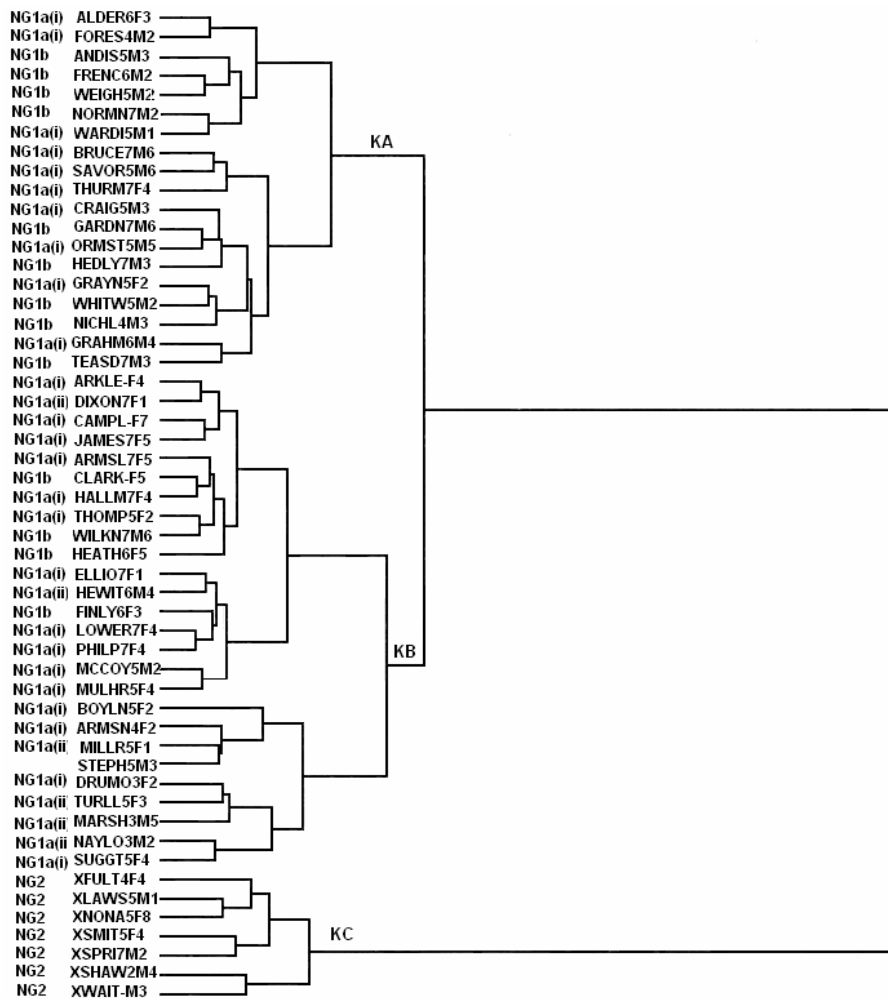


Figure 12: Cluster tree for diphthong data from Jones-Sargent (1983: 198)

To facilitate comparison, we have inserted to the left of Jones-Sargent's labels at the leaves of the tree our own cluster labels from Figure 10a showing, for each speaker, the cluster to which that speaker belongs in our own analysis. The sharp distinction we found between the Newcastle and Gateshead speakers is evident here., as well as in Jones-Sargent' s analyses of the TLS monophthong and consonant groups. Apart from this, however, it is difficult to see anything but a random match between our results and Jones-Sargent's.

As regards the correlation between phonetic clusters and social factors, Jones-Sargent's results are very different from our own: her conclusion is 'that there is no simple relationship between the social classification and this linguistic classification' (1983:249). A full explication of this disparity in results would require engagement with the details of Jones-Sargent's methodology

for social classification of the TLS speakers (1983, chs. 5 & 7), which is not possible within the space constraints on this discussion. It must therefore suffice to note that she clustered speakers on 38 social variables using a methodology closely analogous to that used for the phonetic data, and then attempted to correlate the social and phonetic cluster results. We took the far less complex approach of manually identifying the social variables for which a correlation with the phonetic cluster results could be demonstrated or seemed likely.

More generally on the relationship between phonetic clusters and social factors, previous analysis has found a consistent correlation between a number of social variables and phonetic / phonological variation in Tyneside English. Milroy *et al* 1994, Docherty *et al* 1997 and Watt & Milroy 1999, analyze the social distribution of variants of a small number of linguistic variables (/p/, /t/ and /k/, and the FACE, GOAT and NURSE vowels), and they all find that gender, in particular, plays a central role in determining the distribution of linguistic variables in the dialect. So important is this social factor that Watt & Milroy (1999:42) suggest that that “ Differentiation by gender ... seems in fact to be tantamount to a sociolinguistic priority” in their data. The distribution of male and female speakers in the present analysis suggests that gender is a centrally important factor in determining the distribution of a wide range of linguistic variants, of which the small number of variables previously examined are only a small part. Similarly, Milroy *et al* 1994, Docherty *et al* 1997 and Watt & Milroy 1999 all find that social class and age have an important effect on the distribution of the linguistic variants that they examine. For example, Watt & Milroy (1999) found that highly localized, traditional pronunciations of the FACE, GOAT and NURSE vowels are most characteristic of older working class speakers, whilst “ non-traditional supra-local” variants (p.40) are most characteristic of younger middle class speakers.

5. Conclusion

The aim of the research reported in this paper was to generate hypotheses about phonetic variation among speakers and speaker groups in the NECTE corpus, and how this variation correlates with social factors, using exploratory multivariate analysis and cluster analysis more particularly. The result was a clearly-defined classification of speakers on the basis of their phonetic usage that has a strong correlation with the speakers' social characteristics, and generally agrees with existing work on Tyneside English.

Future work based on this result will consider the following:

a) The cluster analysis presented above does not give the 'true' analysis of the data, as already noted: other distance measure / cluster definition combinations in hierarchical analysis as well as completely different analytical methods such as self-organizing maps (Kohonen 2001) can and do yield different results (Jones & Moisl 2005). The aim is therefore to try a variety of other types of analytical method to determine the degree to which they agree with one another.

b) The above analysis says *that* the NECTE speakers fall into particular clusters, but not *why*. We aim to establish this by examining the clusters with a view to identifying the phonetic variables that are most important in determining the cluster structure.

6. References

Baeza-Yeates, R., Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, Addison Wesley.

Belew, R. (2000) *Finding Out About: a cognitive perspective on search engine technology and the WWW*, Cambridge University Press.

Corrigan, K., Moisl, H., Beal, J. (2005) *The Newcastle Electronic Corpus of Tyneside English*, <http://www.ncl.ac.uk/necte/>.

Docherty, G. J. and Foulkes, P. (1999) 'Sociophonetic variation in 'glottals' in Newcastle English', Proceedings of the 14th International Congress of Phonetic Sciences, pp. 1037-1040, University of California, Berkeley.

Docherty, G. J., Foulkes, P., Milroy, J., Milroy, L. & Walshaw, D. (1997) ' Descriptive adequacy in phonology: a variationist perspective" , Journal of Linguistics, 33, 275-310.

Everitt, B. (2001), *Cluster Analysis*, 4th ed. London: Arnold.

Hair, J., Anderson, R., Tatham, R., Black, W. (1998) *Multivariate Data Analysis*, 5th ed., Prentice-Hall.

Jones, V. (1985) ' Tyneside syntax: A presentation of some data from the Tyneside Linguistic Survey' , in Viereck, W. (ed.) *Focus on England and Wales*, 163-177. Amsterdam: John Benjamins.

Jones, V. and Moisl, H. (2005) ' Cluster Analysis of the *Newcastle Electronic Corpus of Tyneside English*: A Comparison of Methods' , *Web X: A Decade of the World Wide Web, Proceedings of the Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, Athens, Georgia. *Literary and Linguistic Computing*, 20: 1-22.

Jones-Sargent, V. (1983) ' Tyne Bytes. A computerised sociolinguistic study of Tyneside' , *Bamberger Beitrage zur englischen Sprachwissenschaft*, Peter Lang.

Kohonen, T. (2001) *Self-organizing maps*, 3rd ed., Springer.

Milroy, J. et al. (1994) 'Local and supra-local change in British English: the case of glottalisation', *English World-Wide*, 15 (1): 1-32.

Milroy, L., Milroy, J., Docherty, G.J., Foulkes, P. & Walshaw, D. 'Phonological variation and change in contemporary English: evidence from Newcastle-upon-Tyne and Derby', in Condž Silvestre, J.C., & Hernandez-Compoy, J.M. (eds.) *Variation and Linguistic Change in English*, pp.35-46. Cuadernos de Filolog' a Ingelsa.

Pellowe, J., Nixon, G., Strang, B., McNeany, V (1972) ' A dynamic modelling of linguistic variation: the urban (Tyneside) Linguistic Survey' , *Lingua* 30, 1-30.

Pellowe, J., & Jones, V. (1978) ' On intonational variety in Tyneside speech' , in Trudgill, P. (ed.) *Sociolinguistic Patterns of British English*, pp.101-121. London: Arnold.

Strang, B. (1968) ' The Tyneside Linguistic Survey' . *Zeitschrift für Mundartforschung*, NF 4 (Verhandlungen des Zweiten Internationalen Dialektologenkongresses), pp.788-794. Wiesbaden: Franz Steiner Verlag.

Tabachnik, B., Fidell, L. (2001), *Using Multivariate Statistics*, 4th ed., Allyn & Bacon.

Watt, D. & Milroy, L. (1999) " Patterns of variation and change in three Newcastle vowels: is this dialect levelling?" . In Foulkes, P. and Docherty, G. (Eds.), *Urban voices: accent studies in the British Isles*, 25-47. London: Arnold.